Bioregions in marine environments: Combining Biological and Environmental Data for Management and Scientific Understanding

Woolley, Skipton

2020-01

Woolley , S , Bax , N , Currie , J , Dunn , D , Hansen , C , Hill , N , O'Hara , T , Ovaskainen , O , Sayre , R , Vanhatalo , J & Dunstan , P 2020 , ' Bioregions in marine environments: Combining Biological and Environmental Data for Management and Scientific Understanding ' , BioScience , vol. 70 , no. 1 , pp. 48-59 . https://doi.org/10.1093/biosci/biz133

http://hdl.handle.net/10138/311327
https://doi.org/10.1093/biosci/biz133

acceptedVersion

OXFORD UNIVERSITY PRESS | BioScience

Bioregions in marine environments: Combining Biological and Environmental Data for Management and Scientific Understanding

Journal:	*BioScience*
Manuscript ID	BIOS-19-0020.R2
Manuscript Type:	Overview Article
Date Submitted by the Author:	n/a
Complete List of Authors:	Woolley, Skipton; CSIRO, Oceans and Atmosphere Foster, Scott; CSIRO, Data61 Bax, Nicolas; University of Tasmania, IMAS Currie, Jock; Nelson Mandela University, Institute for Coastal and Marine Research Dunn, Daniel; Duke University, Marine Geospatial Ecology Lab Hansen, Cecilie; Institute of Marine Research, Ecosystem Modelling Hill, Nicole; University of Tasmania Institute for Marine and Antarctic Studies, biodiversity modelling O'Hara, Timothy; Museums Victoria, Sciences Department Ovaskainen, Otso; Helsingin Yliopisto, Organismal and Evolutionary Biology Research Progamme Sayre, Roger; US Geological Survey, Land Change Science Program Vanhatalo, Jarno; Helsingin Yliopisto, Department of Mathematics and Statistics Dunstan, Piers; CSIRO, Oceans and Atmosphere
Key words:	biogeography, community ecology, statistics, marine biology
Abstract:	Bioregions are important tools for understanding and managing natural resources. Bioregions should describe where relatively homogenous assemblages of species, enabling managers to better regulate activities that might affect these assemblages. Many existing bioregionalisation approaches, which rely on expert derived, delphic comparisons or environmental surrogates do not explicitly include observed biological data in such analyses. We highlight that for bioregionalisations to be useful and reliable for systems scientists and managers, bioregionalisations need to be based on biological data, include an easily understood assessment of uncertainty, preferably in a spatial format matching the bioregions, and be scientifically transparent and reproducible. Statistical models provide a scientifically robust, transparent and interpretable approach for ensuring that bioregions are formed based on observed biological and physical data. Using statistically-derived bioregions provides a repeatable framework for the spatial representation of biodiversity at multiple spatial scales. This results in better informed management decisions and biodiversity conservation outcomes.

Bioregions in marine environments: Combining Biological and Environmental Data for Management and Scientific Understanding

Skipton N.C Woolley[1], Scott D Foster[2], Nicholas J Bax[1,3], Jock C Currie[4], Daniel C. Dunn[5], Cecilie Hansen[6], Nicole Hill[3], Timothy D O'Hara[7], Otso Ovaskainen[8,9], Roger Sayre[10], Jarno P Vanhatalo[8,11] & Piers K Dunstan[1]

[1]*Oceans and Atmospheres, CSIRO, Hobart, AUS.*
[2]*Data61, CSIRO, Hobart, AUS.*
[3] *Institute for Marine and Antarctic Studies, University of Tasmania, Hobart, AUS*
[4] *Nelson Mandela University & South African National Biodiversity Institute, Cape Town, South Africa*
[5] *Marine Geospatial Ecology Lab, Duke University, Durham, NC, USA*
[6] *Institute of Marine Research, Bergen, Norway*
[7] *Museums Victoria, GPO Box 666, Melbourne, VIC 3001, Australia*
[8] *Organismal and Evolutionary Biology Research Progamme, P.O. Box 65, 00014 University of Helsinki, Finland.*
[9] *Centre for Biodiversity Dynamics, Department of Biology, Norwegian University of Science and Technology, N-7491 Trondheim, Norway.*
[10]*Land Change Science Program, U.S. Geological Survey, Reston, VA, USA*
[11] *Department of Mathematics and Statistics, University of Helsinki, Finland.*

Abstract

Bioregions are important tools for understanding and managing natural resources. Bioregions should describe where relatively homogenous assemblages of species, enabling managers to better regulate activities that might affect these assemblages. Many existing bioregionalisation approaches, which rely on expert derived, delphic comparisons or environmental surrogates do not explicitly include observed biological data in such analyses. We highlight that for bioregionalisations to be useful and reliable for systems scientists and managers, bioregionalisations need to be based on biological data, include an easily understood assessment of uncertainty, preferably in a spatial format matching the bioregions, and be scientifically transparent and reproducible. Statistical models provide a scientifically robust, transparent and interpretable approach for ensuring that bioregions are formed based on observed biological and physical data. Using statistically-derived bioregions provides a repeatable framework for the spatial representation of biodiversity at multiple spatial scales. This results in better informed management decisions and biodiversity conservation outcomes.

Introduction

The distribution of species is a function of many controlling influences operating at a diversity of scales, including environmental heterogeneity and stability in space and time (Rohde 2007), genetic and evolutionary history (Webb et al. 2002), intra- and inter-specific species interactions such as predation, competition and facilitation (Polechová & Barton 2005), dispersal dynamics (Ronce 2007), and environmental disturbances (Sheil 2016). Irrespective of history, the present patterns can be organised spatially creating a biogeography (Last et al. 2010; Ebach & Parenti 2015). With knowledge of where all species exist, scientists would be in a better position to understand why species are distributed as they are, a fundamental line of biogeographic inquiry. Moreover, managers would be in a better position to manage species and their assemblages for a variety of applications, including the conservation and sustainable use of biodiversity and ecosystems, and their associated goods and services.

The quantity and quality of observations of data required to precisely understand where all species are located is impractical to achieve. This is particularly so in ecosystems that are vast and inaccessible, such as our focus: the marine ecosystem. All individual species cannot be surveyed, and, even for well-studied species, complete knowledge on their distributions remains highly uncertain. From a scientific perspective, knowledge of the distribution of species is still fundamentally lacking, despite long-term, ongoing efforts to compile observational datasets for a broad range of taxonomic groups (e.g. Ocean Biogeographic Information System; Grassle 2000). Consequently, to better understand how species are placed within their environment, tools like species distribution models are used to describe their distributions (Guisan & Zimmermann 2000). When considering a few species, the use of individual species distribution models is a logical approach to describe their distributions. However, with many species in a geographical region, researchers may want to move beyond individual species' distributions and better understand the distributions of species assemblages, communities, ecosystems and bioregions (Fig 1a; Ferrier & Guisan 2006; Warton et al. 2015). To do this, scientists often engage in biogeographic classification, otherwise known as bioregionalisation, ecoregionalisation, zoogeographic classification, and ecological mapping (Ebach & Parenti 2015). Here we consider that bioregionalisations are a biological and physical partitioning of geographic space based on the spatial distribution of multiple species, communities, ecosystems, or other biological characteristics. This description shares many concepts with approaches such as vegetation classification, ecosystem characterisation, ecoregions and fisheries regions (Begg et al. 1999).

Bioregions are a simplification (a model) of the true distribution of multiple species that share a similar ecological and abiotic preference and sometimes an evolutionary history. Bioregion maps may be useful for managing multiple different human activities in a region because they simplify complex information into a form that humans are inherently good at understanding (May 1976). Bioregions should define the key physical and biological attributes of a region, provide a simplified understanding of the ecosystems and can be an effective way to compare geographic differences in species composition from local to global scales (Fig. 1b). Bioregions can help contextualise spatial management in a framework that is transparent for decision-making. Decision makers often need to know the spatial extent of biological diversity in order to assess a management action in the context of the biological component of interest (Fig. 1c). Bioregional classifications provide the fundamental building blocks to inform the most appropriate management tools for any geographic area (CBD 2010). Managers might want to assess the impact of human activity (Leaper et al. 2012), gauge the representativeness of a protected area system (Brunckhorst & Bridgewater 1995), or establish representative monitoring programs (Hutchings et al. 2009), within and across bioregions.

Herein, we advocate for the continued development of statistical bioregions to increase scientific understanding of the distribution of biodiversity and to support resource management. We identify the desired characteristics of bioregions, emphasising the importance of appropriate statistical methods in their derivation. We provide a case study in the marine environment to demonstrate one example of how a statistical bioregionalistion can be conducted. As implementation of management based on biogeographic classification continues to be developed, there is a need for rigorous, transparent and well-accepted statistical biogeographic characterisations to deliver improved management tools to support sustainable use and conservation at local, national and global levels.

The current state of marine bioregionalisation approaches

Bioregionalisations often rely heavily on physical, spatial or biological surrogates to describe the distribution of more complex assemblages, communities, ecosystems or bioregions. This approach has been implemented with both expert knowledge (UNESCO 2009) and statistical modelling (Reygondeau et al. 2017; Sayre et al. 2017). These approaches are useful across broad geographic regions where equivalent biological data are lacking or too sparse to inform reliable biological models (Beier & Albuquerque 2015), if their uncertainty is recognised and presented. Global marine bioregional maps have been produced, which ecologically partition the planet based on only abiotic characteristics (environmental drivers) rather than in combination with biotic distributions

(UNESCO 2009; Longhurst 2010; Sayre et al. 2017). These maps attempt to depict ecological zonations based on environmental variation and are often labelled 'ecoregions'. Despite the term 'ecoregions', we stress that these are not ecological, because there is no explicit link to biological data. They do, however, provide a useful partitioning of environment, which might be better coined as 'enviro-regions'. An example of this is the Global Open Oceans and Deep Seabed (GOODS; UNESCO 2009) biogeographic classification that assumes that ocean basins delineate species. However, at least for some taxa this assumption, which seems quite plausible at face-value, is not supported by more recent research (O'Hara et al. 2011).

Recent work by Sayre *et al.,* (2017) provides an example of 'enviro-regions' – regions based on physical data. This physical regionalisation is both broad-scale (global) and relatively fine-resolution (1/4° across the globe). Compared to bioregions, 'enviro-regions' are relatively easy to create as physical data are usually more accessible and comprehensive. These environmental regionalisations can correlate with variation in biotic distributions and are assumed to be representative of biotic distributions. However, some studies have demonstrated less partitioning of geographic space with the inclusion of biological data (Woolley et al. 2013), suggesting that species might persist at broader ranges of environmental variation than the variation generated from physical data alone. Other studies have shown the danger of over-prediction when using physical data alone, without a better understanding of the biology of the species or community, important physical environmental variables may be lacking from the analysis (Anderson et al. 2016). It follows that addition of biological and ecological information can improve delineation of ecological patterns, through improved accuracy, granularity and reduced bias (Warton et al. 2015).

Biological information is often incorporated into bioregionalisations as expert-derived products (Spalding et al. 2007; UNESCO 2009; Longhurst 2010) and in such form rarely includes estimates of uncertainty (Robinson et al. 2017). In such cases, bioregion boundaries are outlined by experts (humans), often as part of a committee, and likely influenced by anthropogenic requirements (e.g. fisheries regions and stocks; Begg et al. 1999; Longhurst 2010). Good examples of this are fisheries areas that reflect governance boundaries and consequently might not accurately describe the distributions of species or ecosystems of interest (Department of the Environment and Heritage, 2006). Expert-derived bioregionalisations are typically quite coarse resolution and often broad in spatial extent (Ekman 1953). While expert information is often easily communicated and is applicable in situations where data are inadequate, it may not be objective or reproducible.

An alternative approach (which could be based on physical or expert-derived data), is to use *statistical* models that estimate the distribution and content of bioregions based on biological and physical data. Increasingly improved global datasets that incorporate remotely sensed information (including both satellites and autonomous platforms to provide information on the surface and sub-surface physical properties) are reducing the need to rely on physical surrogates or expert-derived processes when generating bioregions. Much of the difficulty in producing a bioregion stems from relating (dense) interpolated physical data layers to (sparser) biological data.

What data can inform statistical bioregions?

The generation of bioregions requires data. The three main types of data for bioregionalisation are biological data, physical data and expert-derived knowledge (which itself is usually a mental model based on the first two data types). The volume and variety of biological and physical datasets are increasing, and in many areas, we have reached a key point in time where bioregionalisation can now evolve towards data-driven analyses and products based on observed data and expressed with accompanying measures of uncertainty.

Physical data have many desirable features which make them good datasets for informing bioregionalisation; notably good spatial coverage. 'Physical data' is a term we use here to represent many types of environmental, abiotic, geomorphic or spatial data used to inform the classification of biological data and represent the physical world. The main sources of synoptic physical data include remotely sensed data, model outputs and interpolations of *in situ* physical data. At broad spatial scales, most remotely sensed data for marine habitats comes from satellites, and *in situ* physical data which has been collected at discrete locations in time, which is modelled and then predicted across space based on these observations. These datasets can come with inherent biases which are often overlooked in broad-scale modelling (Foster et al. 2012). Despite this, physical data can be used to inform the distribution of biological data. Like approaches such as species distribution models, we can generate bioregions based on model relationships between the physical and biological data (Foster et al. 2013).

Biological data comes in many forms, such as genes, traits, populations, species and higher taxonomic units. For our purposes, we will focus on biological data that can be readily incorporated in statistical models to build bioregions or components of bioregions. This largely constrains us to the use of observational data about species (and/or Operational Taxonomic Units) in time and space. These observations can be grouped into two broad categories; data from scientific surveys

and ad-hoc datasets (Graham et al. 2004). Scientific survey data tends to be more systematic and are usually more suitable for scientific endeavours. They include information on the amount of biological material (presence-absence, abundance or biomass) at relatively fine taxonomic resolutions and can include additional biological data like genetic information and/or trait information. The short-coming of survey data in marine environments is that they usually focus on relatively small geographical regions, however there are exceptions to this rule (Edgar & Stuart-Smith 2014). Ad-hoc datasets come from a variety of sources, including museum records and citizen science programs. Generally, they are collected without a rigorous scientific survey design (Warton & Shepherd 2010). Often the location where species were observed is recorded, but corresponding information on absences, survey effort and observation methods are generally lacking. These data are widely referred to as 'presence-only' data and are included in biodiversity databases such as Ocean Biogeographic Information System (OBIS; Grassle 2000;). Presence-only data obtained from biogeographic databases are widely used for modelling broad scale biodiversity patterns. This is because they have the greatest spatial coverage at regional and global scales, however the lack of an appropriate sampling design, and the frequent lack of recorded absences, means that they should be treated with care in statistical biogeographic models, or indeed any inference (Beck et al. 2014).

Expert opinion has had a prominent role in the development of bioregions (Ekman 1953). This is because a major limiting factor to developing many broad-scale bioregionalisations has been the lack of biological (and to a lesser extent, physical) data. Therefore, past bioregionalisation efforts have heavily relied on expert elicitation from taxonomists, marine ecologists, biogeographers and stake-holders to delineate important biogeographic regions based on the current status of literature and local knowledge. Expert opinion is still likely to play an important role in bioregional analyses, as it contains implicit information on a region and how species might be distributed within it, as well as an understanding of the biases associated with different data types or surveys. One major issue with expert-based bioregionalisations is reproducibility and assessing the uncertainty in predictions, as expert knowledge is a synthesis of mental, rather than statistical, models. However, there are promising methods that can explicitly include expert knowledge as prior information into statistical models, which we discuss below.

Developing statistical bioregions

A bioregion can be defined as a geographic region with some relatively constant biological characteristics, while the biology across different bioregions are relatively different (Brunckhorst & Bridgewater 1995). This definition is intuitively appealing in its logic, but it is not specific enough

to guide formal data analysis. To formalise it, we need to define characteristics of a bioregion and specify how they should reflect in biological data. A formal definition of bioregions enables their description in the context of their spatial domain and their relationships to physical data, which can be used as explanatory variables to inform a model. Under all appropriate statistical approaches, we suggest a useful characteristic of a bioregion is an area in which the community composition (the set of species attributes, such as their abundances) is approximately constant. Different bioregions are characterised by different community composition and their respective relationships to the physical data (Fig. 1). A similar formal definition was introduced by Ter Braak et al. (2003) and Foster et al. (2013) using presence and absence data, and expanded to count and biomass data of each species to include a constant abundance within each bioregion (Foster et al. 2017). However, such a definition requires careful implementation when the data arise from samples that have different areal or temporal units of measurement. In such cases, the scale of the data is different and must be adjusted for during the analysis – generally using an offset in the model (e.g. Foster et al. 2017). However, using quantities such as probability of occurrence needs to be interpreted with information on how the data were collected to effectively describe the probability with reference to the sampling unit (Warton & Shepherd 2010). Like most classification approaches we assume that once bioregions are defined, the species composition remains constant per bioregion.

Currently, there are many statistical approaches available to classify biological data in to bioregions. However, the choice of which approach to take will often be dictated by the type of data available and the inferences the researchers wish to make. We suggest a useful delineation of possible approaches into the following four categories (like those suggested by Ferrier & Guisan 2006)

1. **Predict First, then Group:** A two-step procedure that involves predicting the value of each species at a grid of locations and then clustering those predictions. The environmental conditions are incorporated in the first step through species distribution models (Guisan & Zimmermann 2000), which output species prediction maps. The set of individual species maps are used as inputs into a spatial clustering analysis in a second step. There are multiple model choices available for each step of this analysis. For the prediction step, any kind of species distribution modelling procedure appropriate for the input data could be used (Guisan & Zimmermann 2000). For the clustering step, the analytical method should ideally have methods that will help inform the number of bioregions/clusters (e.g. k-means clustering or model-based clustering; Fraley & Raftery 2002).

2. **Jointly Predict, then Group:** This is an extension of the previous method, where recent developments in joint species distribution modelling (Thorson et al. 2016; Ovaskainen et al. 2017; Vanhatalo et al. 2018) enable the joint estimation of multiple species and their interspecific correlations (Thorson et al. 2016; Ovaskainen et al. 2017; Vanhatalo et al. 2018). Predictions from the multispecies JSDM are passed to an appropriate clustering method to group species into regions. This remains a two-step procedure for delineating bioregions and does not explicitly aim for spatially contiguous regions (Ovaskainen et al. 2017).

3. **Group First, and then Predict:** Another two-step approach involves first clustering biological data alone, and then predicting the clusters into unsampled locations using a variant of a 'species' distribution model (Miller & Franklin 2002; Ohmann & Gregory 2002; Vogiatzakis & Griffiths 2006). These are similar steps to the previous methods but are performed in the reverse order. Like before, there are multiple choices for appropriate methods to be used in each step.

4. **Analyse Simultaneously:** Perform both clustering and spatial predictions within a single model that defines the assumptions/requirements of a bioregion, propagates uncertainty throughout the process and appropriately handles the multivariate spatial data (Ter Braak et al. 2003; Foster et al. 2013, 2017; Valle et al. 2014).

Each method has its positive and negative attributes, but some are inappropriate for certain situations. The choice among them will depend on the kind of results required and the kind of data available. As we describe above, there are two main sources of biological data, those that come from scientific surveys and those collected in an ad-hoc manner. Currently, many of these methods have been described and build on scientific survey datasets, where the collection of biological data is relatively consistent between observations. This means that for many of these approaches systematic sampling is required to generate robust bioregional outputs. Currently, the 'Predict First, then Group' approach is one of the few approaches which can assemble bioregions based on ad-hoc data. This approach allows for the development and prediction of species-specific presence-only models (Warton & Shepherd 2010). These single species predictions can then be classified into bioregions based on species which have similar ad-hoc collection sightings. At a broad geographical extent, the use of the 'Predict First, then Group' approach is useful with reference to presence-only datasets and methods account for issues associated with variability in occurrence records (Warton & Shepherd 2010). Broad scale bioregionalisations have been achieved using multiple presence-only species distribution models, which are then clustered to provide insight into major biogeographical configuration (O'Hara et al. 2011; El-Gabbas & Dormann 2018). These approaches still suffer from

a range of issues, especially related to observational biases in occurrence record data. Correcting for observational and taxonomic biases in broad scale occurrence data is an active area of statistical research (Renner et al. 2015).

Some commonly used 'Predict First, then Group' approaches are based on biological distances (e.g. Bray-Curtis dissimilarity). These are fed into regression type models to predict biological dissimilarities in unobserved regions, based on site-pair differences in the physical data (Ferrier & Guisan 2006), and subsequently classified into similar ecological regions or clusters. This approach fails to capture several key statistical principles making it inappropriate as a statistical method for bioregional classification: Firstly, they do not model the observed data, but rather an algorithmic abstraction of it (dissimilarities), which means that concepts like mean-variance relationships are often violated (Warton et al. 2012). Secondly, the model likelihoods are often inappropriately specified as it is based on models for single observations, not pairs of observations; so derived metrics from the likelihood such as information criteria and deviance are unreliable (Warton et al. 2015). Thirdly, they typically ignore uncertainty or are unable to compute it directly (Woolley et al. 2016).

Recent development of joint species distribution models has seen their application in specific ecological and biogeographic contexts. Joint species distribution models (JSDMs) are a powerful extension to the 'Predict First, then Group' approach, because they jointly estimate the covariance between species, which improves prediction and provides insight into how species are related and structured (Hui et al. 2013; Thorson et al. 2016). They require subsequent clustering on the species predictions, which if done with appropriate clustering methods, should produce reliable results. While these powerful approaches are at the cutting edge of ecological statistics, they currently fail to propagate the uncertainty from the species level predictions through to the bioregional classification step (Ovaskainen et al. 2017). As a result, their predicted bioregional classifications lacks an estimate of uncertainty in the bioregional predictions, however this information can be obtained at the species level (Warton et al. 2015).

The 'Group First, then Predict' method suffers from many of the criticisms, and benefits from similar strengths, to those of the 'Predict First, then Group' method. A positive, compared to 'Predict First, then Group', is that the number of prediction models is greatly reduced. This enables the analyst to really focus on fitting good models and diagnosing them well (Miller & Franklin

2002; Vogiatzakis & Griffiths 2006). Unlike the 'Predict First, then Group' method, grouping first currently restricts completely the use of ad-hoc data as methods to cluster only presences are undeveloped. This severely limits the breadth of applications it is available for. Lastly, both the 'Group First, then Predict' and the 'Predict First, then Group' methods typically fail to propagate uncertainty from the data through to the final bioregional classification.

Examples of the 'Analyse Simultaneously' bioregional methods have recently emerged (Dunstan et al. 2011; Foster et al. 2013, 2017). These approaches build upon the concepts of modelling physical and biological data together, but do the prediction and clustering within a single model. These approaches model observed data directly and transfer the variance of the data all the way through to final bioregional prediction (Woolley et al. 2013; Hill et al. 2017).

We argue for the purposes of bioregionalisation using a model which is designed specifically for estimating bioregions should be used. The 'Analyse Simultaneously' approaches can account for inter-dependencies between biological and physical data when estimating bioregional classifications (Foster et al. 2013, 2017). Researchers might achieve what they consider ecologically informative regionalisation using any of these four approaches, but must be aware of the information lost at each modelling step in a bioregional analysis (Hill et al. In Prep).

Case study

To illustrate how one might implement a statistical bioregionalisation, we present a bioregionalisation of fish on the North-West shelf of Australia. The analysis was performed using an extension of the Regions of Common Profiles (RCP) model (Foster et al. 2013) that allows for spatial coherency (Vanhatalo et al. In Review) and is an example of an 'Analyse Simultaneously' method. There are several important decisions which need to be considered when developing these approaches. Firstly, from a biological perspective we need to consider the number of species to include in the model and what are the minimum number of observations a species requires to be included. As a rule of thumb, multiple species models such as JSDM and mixture models can handle rarer species compared to single species models (Hui et al. 2013; Norberg et al. 2019). In this case study, the entire dataset consisted of 854 demersal trawls taken in depths of 20 to 450 m from October 1986 to August 1997. Each trawl sampled approximately the same amount of seabed, so no adjustment is necessary for varying sample effort. We based the bioregionalisation on 253 teleost and chondrichtyan species, from a total of 579 species. We chose this subset as of species as

they were observed in a least 15 or more trawls. As a rule of thumb, multiple species models such as JSDM and mixture models can handle rarer species compared to single species models (Hui et al. 2013; Norberg et al. 2019). Species observed in fewer trawls could have been included in the analyses if the distribution of rare (and potentially threatened) species was a management priority. However, the inclusion of these species would likely add additional noise making it harder to extract relevant information.

As per the biological data, the choice of physical data used as covariates in the modelling will have important ramifications for the bioregionalisation produced. Ideally, the covariates used in the model should best describe the environmental and abiotic factors which characterise each species distribution. For our case study, we chose intra-annual standard deviation (SD) of nitrate, intra-annual SD of dissolved oxygen, annual mean of salinity, intra-annual SD of silicate and intra-annual SD of sea surface temperature as physical data to define bioregions (see Foster et al., 2013 for details). Intra-annual variation can be important to ecological systems as it measures the range of environmental conditions that a single location may encounter. In this example, we did not include information on geomorphic data like soft and hard substrate. These types of variables are likely to be important for describing marine species distributions and will help inform species distributions and assemblages which respond strongly to physical features, rather than environmental gradients (like azonal ecosystems in terrestrial environments; Olson et al. 2001). It is quite plausible that different bioregions can exist in the same covariate space. In these instances, this would likely be an effect of missing covariates, which could be added to an analysis (if available) to help differentiate bioregions. Different physical data will tend to operate on different spatial and temporal scales which could have important implications for bioregionalisation and the variation of assemblages (Austin 2002).

The number of groups chosen during the bioregionalisation process can drastically change the bioregional outcomes (Miller 1996). In the RCP approach we estimated the number of groups from the data based on the model likelihood. Using a single step, sites are grouped based on the species composition and their relationship to the physical data. The model likelihood was then used to inform the number of groups based on Bayesian Information Criterion (BIC; Burnham & Anderson 2004). Our model identified four bioregions. Choosing the number of groups which best represents the available data appears to be one of the key advantages of the ‘Analyse Simultaneously’ method. Other approaches can generate similar groupings, but they must be done in a two-step process,

which potentially divorces the link between the biological data and number of groups (Hill et al. In Prep).

In this analysis we kept outputs simple for illustration purposes, but note that this analysis can provide more complex outputs. We present a discrete (or hard) classification by assigning site labels based on the most probable bioregion at that site, even though the probability of each site belonging to each bioregion (RCP group) is estimated. The discrete clustered bioregions are given in Fig. 2b, which suggests that there is a coastal region, an inner continental shelf bioregion, a patchy mid-shelf bioregion, and an outer shelf and slope bioregion. Greater information can be gained by examining the probabilities of each bioregion being present across the same study region (Figs. 2c-f). There appears to be quite high probability for Bioregion 1 throughout the entire shallow and medium-water environment and this overlaps substantially with Bioregions 3 and 4. Conversely, the deep-water bioregion (Bioregion 2) appears to have a sharp boundary where the continental margin descends more steeply. Uncertainty maps are available for the probabilistic prediction, as illustrated for Bioregion 3 (See Fig. 3a-c). There are many spatial locations where the predicted presence of this bioregion has low certainty. This is evidenced by the interval estimates (95% confidence interval), Fig. 3a and Fig. 3c, covering the probabilities between (almost) zero and (almost) one. However, there are locations where the probability is certain. These include locations where the bioregion is not (e.g. in the deeper and shallowest water) and locations where there is high confidence in the prediction.

To understand the predicted biological content of each bioregion, we can inspect its species profile. An example, again for Bioregion 3, is given in Fig. 3d: this Bioregion is represented by a small number of species that are very likely present (probability of observation > 0.5); a moderate number of species that are moderately likely to be found; and many species that are unlikely to be present. Summing these probabilities gives an indication of species richness in the bioregion. In this case, we would expect to encounter approximately 37 species each time a trawl is performed at a site estimated to have high probability (probability of observation > 0.5) of belonging to this bioregion. The species profiles also enable contrasts between bioregions based on their biological content, the profile is the prevalence of every species in each bioregion (we have depicted the profile as a line in Fig. 3, but the identities and their prevalence can be compared across regions; e.g. Hill et al. 2017): if two bioregions share a similar species profile, then they are less different than two bioregions that do not.

Statistical bioregionalisations offer a robust means for identifying, framing and predicting the distribution of biodiversity patterns. The example above shows how quantifying the distributions of multiple species can be distilled into bioregional predictions. These predictions and their associated uncertainties can be assessed against management actions or industrial activities within a bioregion (Fig. 1).

How do statistical bioregions help improve management decisions?

The choice to undertake a bioregionalisation process is often driven by the desire to understand how multiple species are assembled and how best to manage them (Fig. 1a). It is important that information obtained from a bioregionalisation analysis be directly applicable to the current management or scientific question at hand *and* it should be presented with an appropriate level of detail so that it can be understood by those who use it. To us, key considerations in this context include:

(i) *Identify which species are likely to be found in each bioregion (noting that some species may be found in multiple bioregions)*. Understanding the membership of species to each bioregion is a critical step for management because it gives managers the capacity to identify which species will be affected by activities (protective or threatening) in a region. We can see from figure one (Fig. 1b), that many of the species are present across multiple bioregions, but their relative intensity is specific to each region. It is this combination of predicted abundances (or prevalence) for a set of species which represents the community composition present within that bioregion relative to the variation in the physical data (e.g. environmental gradients). For all bioregional approaches, except 'enviro-regionalisation', the species composition of groups can be identified. For all the two-step approaches species in groups can be identified by summarising the observed species' data at classified survey sites, while the 'analyse simultaneously' approach we can estimate the species membership from parameters in the models and the associated uncertainty that each species belongs to a bioregion. Reporting estimates of species density (or prevalence) and the uncertainties associated with those estimates, will further help managers avoid or protect critical areas within a bioregion(s) where key species or assemblages need to be managed (Fig. 1b).

(ii) *Identify which physical data characterise each bioregion.* All approaches should enable the characterisation of physical data used to describe each bioregion (except for expert derived).

Like describing species membership, the two-step approach can summarise the observed environmental data at classified survey site. While the one-step approach can report these characteristics via model parameters (Hill et al. 2017).

(iii) *Identifying the number of bioregions is an important part of any bioregionalisation process.* Choosing the number of bioregions is often driven by the requirements of managers or is chosen to reflect governance boundaries (Department of the Environment and Heritage, 2006). Ideally, the number of bioregions should be informed by the data, the 'Analyse Simultaneously' approach is currently the only approach which can estimate the number of groups with reference to the original data. Having said this, all approaches should perform similarly if the number of bioregions is known (Hill et al. In Prep). When the number of bioregions is unknown, additional information like phylogeny might help inform this step (Ebach & Parenti 2015).

(iv) *Bioregional classification should be undertaken using a transparent analytical process*, so that it is clear to an interested onlooker what was done, why certain decisions were made and what assumptions (ecological and statistical) these decisions reflect. This is a clear advantage of statistical bioregionalisation over expert derived or delphic approaches. Under all statistical bioregional approaches the steps from data, through to analysis, outputs and interpretation can be clearly reported and reproduced based on the data and methods used.

(v) *Bioregional classification should be updatable with the availability of new information*, so that the bioregions can be updated in a coherent and consistent manner when additional data become available. This is clearly an advantage of all statistical bioregionalisation approaches where outputs are derived based on modelled data and clearly reported steps and assumptions in a way that expert-derived products are not.

(vi) *Understanding how uncertainty informs confidence in the location of bioregions*, along with the confidence in the description of physical and biological characteristics within each bioregion (Brown 1998; Fiorentino et al. 2018). Assessments of uncertainty and variance are already standard in many management actions, for example fisheries ecosystem-based management (Koen-Alonso et al. 2019) and are likely to become more important in bioregionalisation decision-making, where economic, social and biodiversity values are often traded to meet competing objectives. Uncertainty helps modellers, ecologists, managers understand how reliable a bioregional classification might be, what its limitations

are (e.g. for widely ranging species) and should be based on the propagation of variance from the data through to the estimated model. It is important to recognise that many management processes (e.g. design of a representative marine reserve network, or an environmental offset program), require stability from bioregional analyses, especially over the time period that policies are being developed. Defining a coherent and consistent process to update bioregionalisations is an important aspect of their value to government.

Future directions for statistical bioregions

Throughout this paper we have advocated the use of data informed statistical bioregionalisations. We acknowledge that a lack of quantitative data has been a limiting factor in many cases, leading to the use of physical data or expert knowledge to characterise bioregions (Reygondeau et al. 2017). The availability of biological occurrence data is escalating rapidly with improved data sharing and collection technologies. However, broad-scale biological datasets frequently contain significant biases and error, such as spatial bias in where occurrences were recorded (near major human populations) and an over representation of rare species (taxonomists are inherently interested in rare species) and gear or observer selectivity (Graham et al. 2004). These present challenges which need to be addressed when developing broad scale bioregional models. In our case study, we used biological data collected in a systematic and consistent way, meaning that abundances (presence and absence) of species was explicitly recorded. If these data had been ad-hoc collections, we currently would not have been able to use RCP model approach to describe bioregions, as the model has not been correctly formulated to handle presence-only occurrence data where absences are not systematically recorded. With a lack of recorded absences, an appropriate model would be a spatial point process (Cressie 1993). Under a situation where we only had presence-only data from ad-hoc surveys, we might choose to use a 'Predict First, then Group' approach where we generate many point process species distribution models independently and then undertake a clustering of the species Poisson point process predictions across the seafloor to generate bioregions (e.g. O'Hara et al. 2011). Future work that can extend presence-only species distribution models (Warton & Shepherd 2010) to multiple species might provide a promising solution. However, as is the case with single-species presence-only approaches, the underlying biases will need to be clearly documented to make users aware of their potential effects in the outputs. Another promising field of statistical model development consists of augmented approaches that combine *ad-hoc* and scientific survey data to generate more robust predictions without having to remove potentially biased *ad-hoc* occurrence records (Fithian et al. 2015; Renner et al. 2015). Such augmented models could possibly be expanded to accommodate multiple species to inform statistical bioregionalisation.

Throughout this manuscript we have largely focused on species as the biological data for bioregionalisation. But it is plausible that these methods could be extended to encapsulate other sources and types of biological data. For example, a similar approach to the RCP model we described in the case study, has been used to understand population genetic structure in stocks of commercially valuable fisheries (Grewe et al. 2015). Similar models could be extended to multiple species, to understand where genetic populations for numerous species are differentiating. For example, the development of eDNA sampling protocols appears to be a promising area where bioregionalisation could be undertaken on Operational Taxonomic Units, to describe the biogeography of important groups such as bacteria and phytoplankton (Rees et al. 2014). Genetic data can also be used to understand the evolutionary processes that shape the distributions of extant species (Webb et al. 2002; Ebach & Parenti 2015) and is an important source of historical information we have largely ignored. The development of new multiples species models (JSDMs) which can explicitly include information on species traits and phylogeny is likely to facilitate new bioregionalisations which incorporate the role of historical processes with reference to observed species in a joint model (Ives & Helmus 2011; Ovaskainen et al. 2017).

Until comprehensive and broad-scale biological datasets preclude its utility, expert knowledge can continue to play a powerful role in bioregional analyses. We acknowledge the importance of expert knowledge as an information source in many cases, but suggest that it be included as informative prior information in a Bayesian framework where possible (Gelman et al. 2013). An effective way to do so might be to elicit information from experts with the aid of probability training (Hosack et al. 2017). Under such an approach, the development of priors based on expert option inform bioregional outputs in low data situations, but as greater volumes of biological data become available, bioregion predictions will increasingly shift towards data driven outcomes.

Conclusion

Statistical bioregions can be used to frame existing ecosystem-based approaches and provide novel insight into how biodiversity is structured. The development of statistical bioregions, and the methodological developments which underpin them, can help build upon existing bioregional classifications by reducing the ambiguity in what a bioregion is, which we have formally defined, along with how this definition can be matched to data. These bioregions add consistency and

reproducibility of classifications over approaches like expert elicitation, whilst making direct assessment of bioregions and the information contained within them derived directly from the data used to generate them. Assessing uncertainty in the quality of the estimated model can improve the decision making based on these bioregionalisations. The implementation of statistical bioregions provides a robust path forward for many scientific problems, but to do so, will require that taxonomists, ecologist, biogeographers, statisticians and stakeholders to work on a common set of problems and integrate their skills into a coherent set of meaningful bioregional products that serve their unique purposes (e.g. scientific inquiry, management, or spatial planning).

Acknowledgments

The 'Statistical Bioregional Workshop' that facilitated this work was funded by the Global Ocean Biodiversity Initiative (GOBI). GOBI is supported by the International Climate Initiative (IKI). The German Federal Ministry for the Environment, Nature Conservation and Nuclear Safety (BMU) supports this initiative on the basis of a decision adopted by the German Bundestag. S.N.C.W was supported by GOBI. C.H was supported by the EAF-Nansen scientific program. N.J.B was supported by the Marine Biodiversity Hub through funding from the Australian Government's National Environmental Science Program. O.O acknowledges funding by the Academy of Finland (grants 273253 and 284601), the Research Council of Norway (SFF-III grant 223257), and by the Jane and Aatos Erkko Foundation (grant to Research Centre for Ecological Change).

References

Anderson OF, Guinotte JM, Rowden AA, Clark MR, Mormede S, Davies AJ, Bowden DA. 2016. Field validation of habitat suitability models for vulnerable marine ecosystems in the South Pacific Ocean: implications for the use of broad-scale models in fisheries management. Ocean & Coastal Management **120**:110–126.

Austin MP. 2002. Spatial prediction of species distribution: an interface between ecological theory and statistical modelling. Ecological Modelling **157**:101–118.

Beck J, Böller M, Erhardt A, Schwanghart W. 2014. Spatial bias in the GBIF database and its effect on modeling species' geographic distributions. Ecological Informatics **19**:10–15.

Begg GA, Friedland KD, Pearce JB. 1999. Stock identification and its role in stock assessment and fisheries management: an overview. Fisheries Research **43**:1–8.

Beier P, Albuquerque FS. 2015. Environmental diversity as a surrogate for species representation. Conservation Biology **29**:1401–1410.

Brown DG. 1998. Mapping historical forest types in Baraga County Michigan, USA as fuzzy sets. Plant Ecology **134**:97–111.

Brunckhorst D, Bridgewater P. 1995. Marine bioregional planning: a strategic framework for identifying marine reserve networks, and planning sustainable use and management. Pages 105–16 Proceedings of the Symposium on Marine Protected Areas and Sustainable Fisheries conducted at the Second International Conference on Science and the Management of Protected Areas.

Burnham KP, Anderson RP. 2004. Multimodel Inference: Understanding AIC and BIC in Model Selection. Sociological Methods & Research **33**:261–304.

CBD. 2010. The strategic plan for biodiversity 2011-2020 and the aichi biodiversity targets. Document UNEP/CBD/COP/DEC/X/2. Secretariat of the Convention on Biological Diversity, Nagoya, Japan.

Cressie N. 1993. Statistics for spatial data. John Wiley & Sons.

Department of the Environment and Heritage, 2006. (n.d.). A guide to The Integrated Marine and Coastal Regionalisation of Australia - version 4.0 June 2006 (IMCRA v4.0). Available from http://www.environment.gov.au/ (accessed February 6, 2018).

Dunstan PK, Foster SD, Darnell R. 2011. Model based grouping of species across environmental gradients. Ecological Modelling **222**:955–963.

Ebach MC, Parenti LR. 2015. The dichotomy of the modern bioregionalization revival. Journal of Biogeography **42**:1801–1808.

Edgar GJ, Stuart-Smith RD. 2014. Systematic global assessment of reef fish communities by the Reef Life Survey program. Scientific Data **1**:140007.

Ekman S. 1953. Zoogeography of the seas. London: Sidgwick & Jackson **953**:415.

El-Gabbas A, Dormann CF. 2018. Improved species-occurrence predictions in data-poor regions: using large-scale data and bias correction with down-weighted Poisson regression and Maxent. Ecography **41**:1161–1172.

Ferrier S, Guisan A. 2006. Spatial modelling of biodiversity at the community level. Journal of Applied Ecology **43**:393–404.

Fiorentino D, Lecours V, Brey T. 2018. On the art of classification in spatial ecology: fuzziness a way to map uncertainty. Frontiers in Ecology and Evolution **6**:231.

Fithian W, Elith J, Hastie T, Keith DA. 2015. Bias correction in species distribution models: pooling survey and collection data for multiple species. Methods in Ecology and Evolution **6**:424–438.

Foster SD, Givens GH, Dornan GJ, Dunstan PK, Darnell R. 2013. Modelling biological regions from multi-species and environmental data. Environmetrics **24**:489–499.

Foster SD, Hill NA, Lyons M. 2017. Ecological grouping of survey sites when sampling artefacts are present. Journal of the Royal Statistical Society: Series C (Applied Statistics) **66**:1031–1047.

Foster SD, Shimadzu H, Darnell R. 2012. Uncertainty in spatially predicted covariates: Is it ignorable? Journal of the Royal Statistical Society. Series C: Applied Statistics **61**:637–652.

Fraley C, Raftery AE. 2002. Model-based clustering, discriminant analysis, and density estimation. J Am Stat Assoc **97**:611–631.

Gelman A, Carlin J, Stern H, Dunson D, Ventari A, Rubin D. 2013. Bayesian Data Analysis. Chapman & Hall/CRC Boca Raton, FL, USA.

Graham C, Ferrier S, Huettman F, Moritz C, Peterson A. 2004. New developments in museum-based informatics and applications in biodiversity analysis. Trends in Ecology & Evolution **19**:497–503.

Grassle JF. 2000. The Ocean Biogeographic Information System (OBIS): an on-line, worldwide atlas for accessing, modeling and mapping marine biological data in a multidimensional geographic context. Oceanography **13**:5–7.

Grewe P, Feutry P, Hill P, Gunasekera R, Schaefer K, Itano D, Fuller D, Foster S, Davies C. 2015. Evidence of discrete yellowfin tuna (Thunnus albacares) populations demands rethink of management for this globally important resource. Scientific reports **5**:16916.

Guisan A, Zimmermann NE. 2000. Predictive habitat distribution models in ecology. Ecological Modelling **135**:147–186.

Hill NA, Foster SD, Woolley S, Dunstan PK, Mckinlay J, Ovaskainen O, Johnson CR. In Prep. Determining marine bioregions: a review by example.

Hill NA, Foster SD, Duhamel G, Welsford D, Koubbi P, Johnson CR. 2017. Model-based mapping of assemblages for ecology and conservation management: A case study of demersal fish on the Kerguelen Plateau. Diversity and Distributions **23**:1216–1230.

Hosack GR, Hayes KR, Barry SC. 2017. Prior elicitation for Bayesian generalised linear models with application to risk control option assessment. Reliability Engineering & System Safety **167**:351–361.

Hui FK, Warton DI, Foster SD, Dunstan PK. 2013. To mix or not to mix: comparing the predictive performance of mixture models vs. separate species distribution models. Ecology **94**:1913–1919.

Hutchings L, Roberts MR, Verheye HM. 2009. Marine environmental monitoring programmes in South Africa: a review. South African Journal of Marine Science **105**:94–102.

Ives AR, Helmus MR. 2011. Generalized linear mixed models for phylogenetic analyses of community structure. Ecological Monographs **81**:511–525.

Koen-Alonso M, Pepin P, Fogarty MJ, Kenny A, Kenchington E. 2019. The Northwest Atlantic Fisheries Organization Roadmap for the development and implementation of an Ecosystem Approach to Fisheries: structure, state of development, and challenges. Marine Policy **100**:342–352.

Last PR, Lyne VD, Williams A, Davies CR, Butler AJ, Yearsley GK. 2010. A hierarchical framework for classifying seabed biodiversity with application to planning and managing Australia's marine biological resources. Biological Conservation **143**:1675–1686.

Leaper R, Dunstan PK, Foster SD, Barrett NJ, Edgar GJ. 2012. Comparing large-scale bioregions and fine-scale community-level biodiversity predictions from subtidal rocky reefs across south-eastern Australia. Journal of applied ecology **49**:851–860.

Longhurst AR. 2010. Ecological geography of the sea. Elsevier.

May RM. 1976. Simple mathematical models with very complicated dynamics. Nature **261**:459.

Miller J, Franklin J. 2002. Modeling the distribution of four vegetation alliances using generalized linear models and classification trees with spatial dependence. Ecological Modelling **157**:227–247.

Miller KR, others. 1996. Balancing the scales: guidelines for increasing biodiversity's chances through bioregional management. World Resources Institute.

Norberg A et al. 2019. A comprehensive evaluation of predictive performance of 33 species distribution models at species and community levels. Ecological Monographs:e01370.

O'Hara TD, Rowden AA, Bax NJ. 2011. A Southern Hemisphere bathyal fauna is distributed in latitudinal bands. Current Biology **21**:226–230.

Ohmann, JL, Gregory, MJ. 2002. Predictive mapping of forest composition and structure with direct gradient analysis and nearest-neighbor imputation in coastal Oregon, USA. Canadian Journal of Forest Research, **32**:725-741.

Olson DM, Dinerstein E, Wikramanayake ED, Burgess ND, Powell GV, Underwood EC, D'amico JA, Itoua I, Strand HE, Morrison JC. 2001. Terrestrial Ecoregions of the World: A New Map of Life on Earth: A new global map of terrestrial ecoregions provides an innovative tool for conserving biodiversity. BioScience **51**:933–938.

Ovaskainen O, Tikhonov G, Norberg A, Guillaume Blanchet F, Duan L, Dunson D, Roslin T, Abrego N. 2017. How to make more out of community data? A conceptual framework and its implementation as models and software. Ecology Letters **20**:561–576.

Polechová J, Barton NH. 2005. Speciation through competition: a critical review. Evolution **59**:1194–1210.

Rees HC, Maddison BC, Middleditch DJ, Patmore JR, Gough KC. 2014. The detection of aquatic animal species using environmental DNA–a review of eDNA as a survey tool in ecology. Journal of Applied Ecology **51**:1450–1459.

Renner IW, Elith J, Baddeley A, Fithian W, Hastie T, Phillips SJ, Popovic G, Warton DI. 2015. Point process models for presence-only analysis. Methods in Ecology and Evolution **6**:366–379.

Reygondeau G, Guieu C, Benedetti F, Irisson J-O, Ayata S-D, Gasparini S, Koubbi P. 2017. Biogeochemical regions of the Mediterranean Sea: an objective multidimensional and multivariate environmental approach. Progress in oceanography **151**:138–148.

Robinson NM, Nelson WA, Costello MJ, Sutherland JE, Lundquist CJ. 2017. A Systematic Review of Marine-Based Species Distribution Models (SDMs) with Recommendations for Best Practice. Frontiers in Marine Science **4**. Available from https://www.frontiersin.org/articles/10.3389/fmars.2017.00421/full (accessed February 5, 2018).

Rohde K. 2007. Latitudinal gradients in species diversity : the search for the primary cause. Oikos **65**:514–527.

Ronce O. 2007. How does it feel to be like a rolling stone? Ten questions about dispersal evolution. Annu. Rev. Ecol. Evol. Syst. **38**:231–253.

Sayre RG, Wright DJ, Breyer SP, Butler KA, Van Graafeiland K, Costello MJ, Harris PT, Goodin KL, Guinotte JM, Basher Z. 2017a. A three-dimensional mapping of the ocean based on environmental data. Oceanography **30**:90–103.

Sheil D. 2016. Disturbance and distributions: avoiding exclusion in a warming world. Ecology and Society **21**.

Spalding MD et al. 2007. Marine Ecoregions of the World: A Bioregionalization of Coastal and Shelf Areas. BioScience **57**:573.

Ter Braak CJ, Hoijtink H, Akkermans W, Verdonschot PF. 2003. Bayesian model-based cluster analysis for predicting macrofaunal communities. Ecological Modelling **160**:235–248.

Thorson JT, Ianelli JN, Larsen EA, Ries L, Scheuerell MD, Szuwalski C, Zipkin EF. 2016. Joint dynamic species distribution models: a tool for community ordination and spatio-temporal monitoring. Global Ecology and Biogeography **25**:1144–1158.

UNESCO. 2009. Global Open Oceans and Deep Seabed (GOODS) - biogeographic classification. IOC Technical Series **84**:84.

Valle D, Baiser B, Woodall CW, Chazdon R. 2014. Decomposing biodiversity data using the Latent Dirichlet Allocation model, a probabilistic multivariate statistical method. Ecology letters **17**:1591–1601.

Vanhatalo J, Foster SD, Hosack GR. In Review. Spatiotemporal Clustering Using Gaussian Processes in a Mixture Model.

Vanhatalo J, Hartmann M, Veneranta L, others. 2018. Additive Multivariate Gaussian Processes for Joint Species Distribution Modeling with Heterogeneous Data. Bayesian Analysis.

Vogiatzakis IN, Griffiths GH. 2006. A GIS-based empirical model for vegetation prediction in Lefka Ori, Crete. Plant ecology **184**:311–323.

Warton DI, Foster SD, De'ath G, Stoklosa J, Dunstan PK. 2015. Model-based thinking for community ecology. Plant Ecology **216**:669–682.

Warton DI, Shepherd LC. 2010. Poisson point process models solve the "pseudo-absence problem" for presence-only data in ecology. Annals of Applied Statistics **4**:1383–1402.

Warton DI, Wright ST, Wang Y. 2012. Distance-based multivariate analyses confound location and dispersion effects. Methods in Ecology and Evolution **3**:89–101.

Webb CO, Ackerly DD, McPeek M a., Donoghue MJ. 2002. Phylogenies and Community Ecology. Annual Review of Ecology and Systematics **33**:475–505.

Woolley SNC, Foster SD, Dunstan PK, O'Hara TD, Wintle BA. 2016. Characterising Uncertainty in Generalised Dissimilarity Modelling. Methods in Ecology and Evolution.

Woolley SNC, McCallum AW, Wilson R, O'Hara TD, Dunstan PK. 2013. Fathom out: Biogeographical subdivision across the Western Australian continental margin - a multispecies modelling approach. Diversity and Distributions **19**:1506–1517.

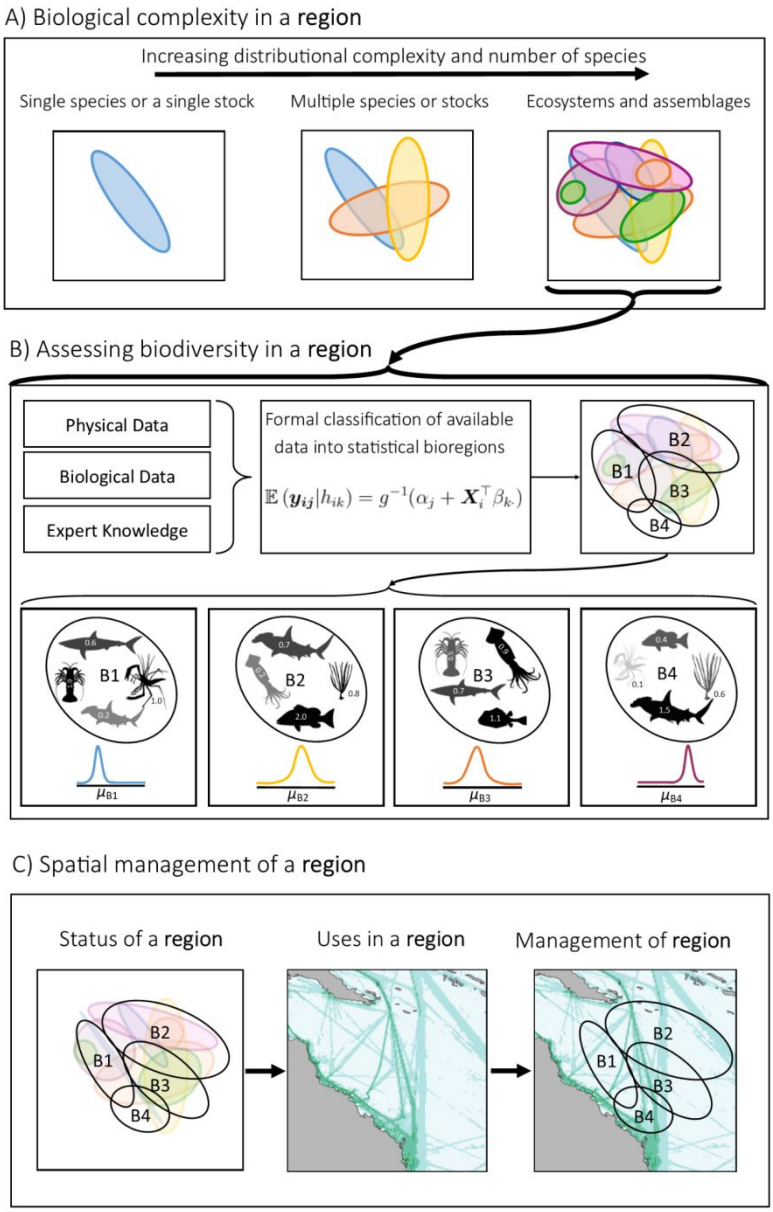


Figure 1. Why do we need bioregions, how can be describe them from data and how can bioregions can be used to manage biodiversity in a region? A) For the management of a single species (or stock) it is often clear how we might model its distribution. But, as the number of species increases it becomes more challenging to model and interpret hundreds to thousands of species. B) Statistical bioregionalisations offer a solution, as they help contextualise and simplify complex ecosystems or species assemblages into units that are understandable and describe the physical and biological characteristics present in each bioregion. Knowledge on the distribution of biological and physical data can be formally incorporated into statistical models and then be used to distil bioregional level predictions or species assemblages. The colour and the numeric value of the species depicted in each bioregion represents the predicted intensity of each species in that region. We can see that some species occur across multiple regions, but their intensities are specific to each bioregion. Although all species will have a predicted intensity within each bioregion, for plotting purposes we have excluded species from the figure where their predicted intensity is effectively zero. C) Once bioregions have been quantified, the distributions of the bioregions and the species therein can be assessed with reference to uses in that region. Bioregions can then help inform decisions about human activities in a region with reference to their impacts on species assemblages within bioregions.

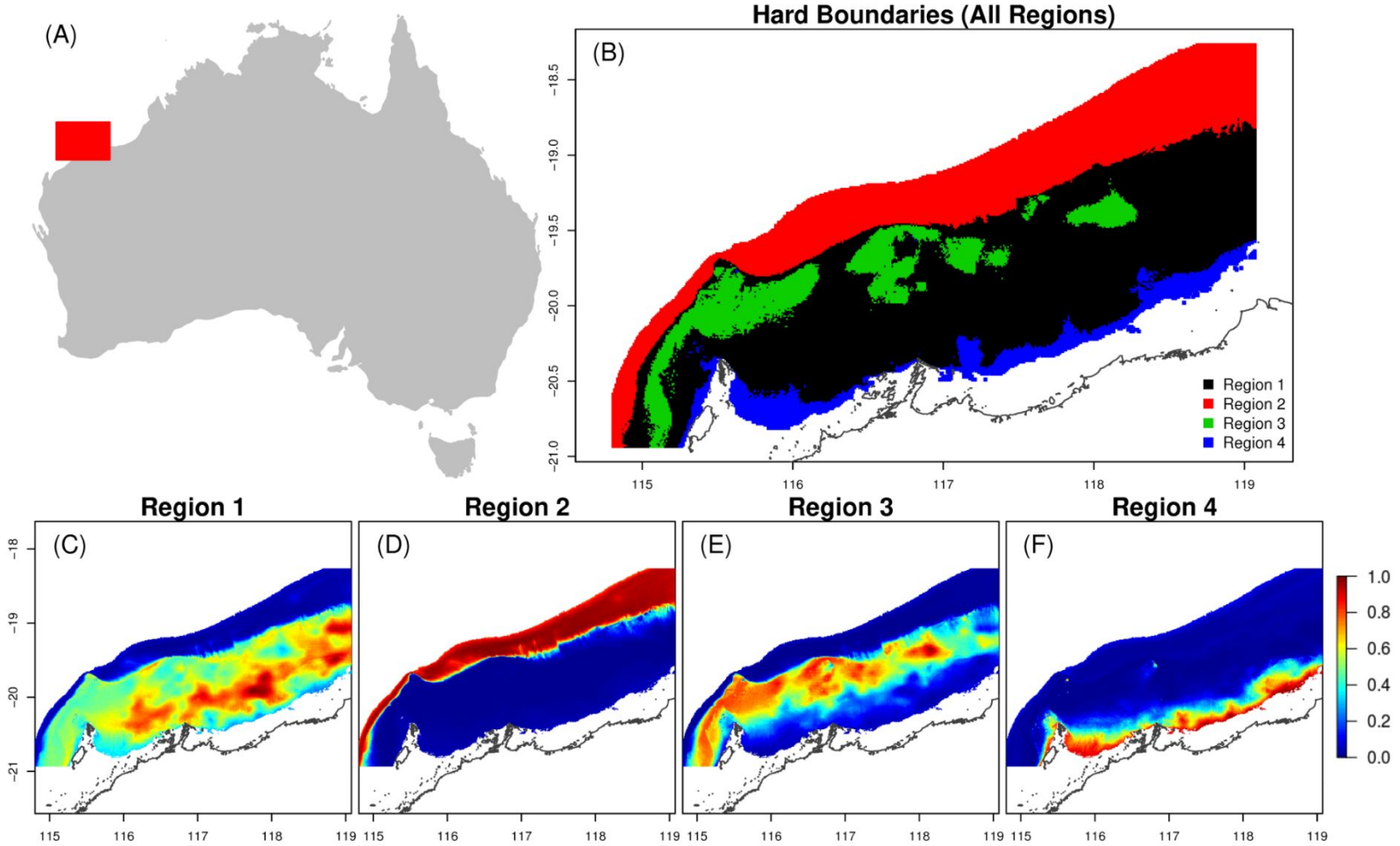
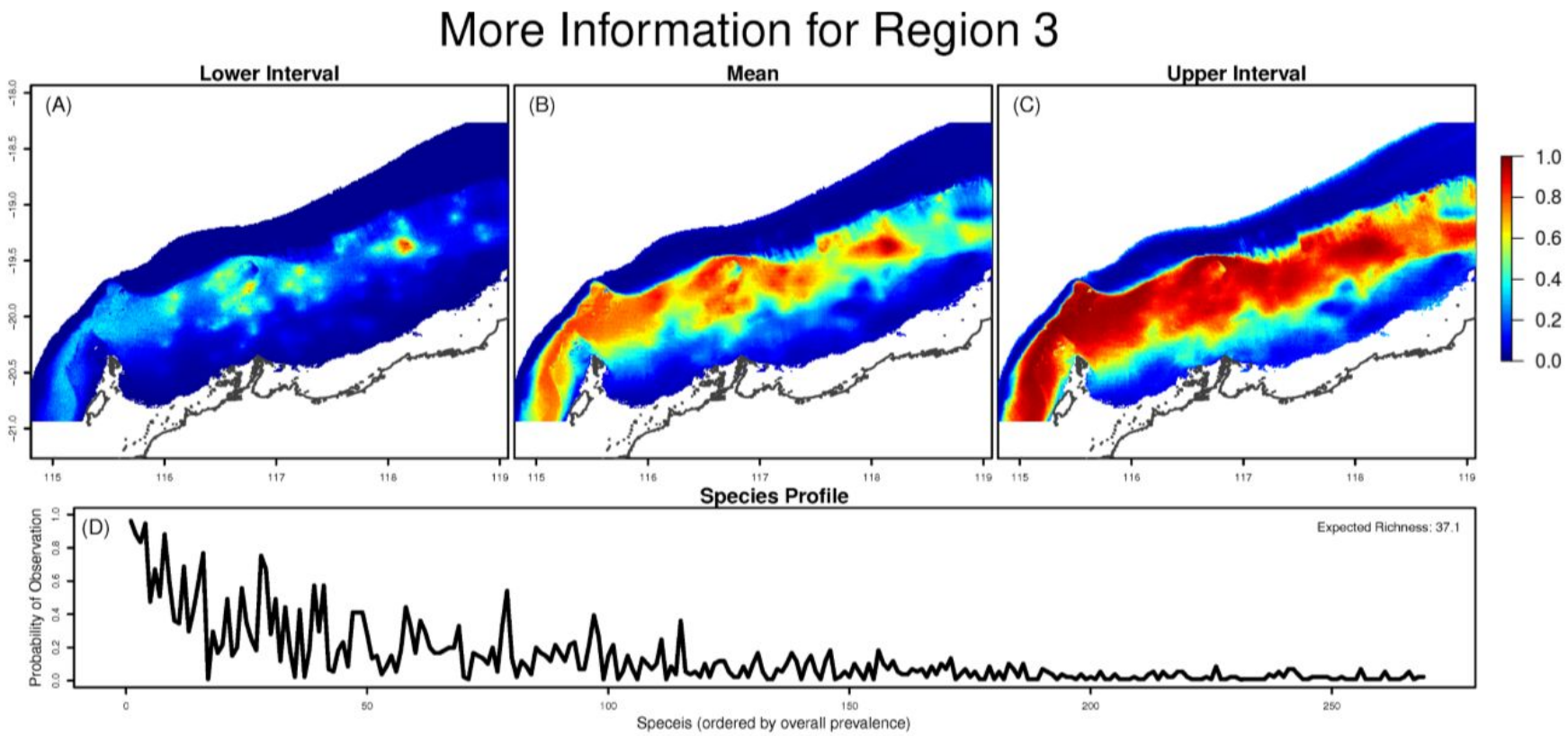


Figure 2. Bioregionalisation of the North-West Shelf area of Australia. (A) Shows where the region is. (B) Shows a set of 4 discrete (hard-clustered) bioregions. (C)-(F) Shows the estimated probability of observing each bioregion in each location. Note that blue colour corresponds to low probability (zero) and red to high (one). Results obtained after applying the regions of common profiles (RCP; Foster et al. 2013), as implemented in Vanhatalo et al (in review).

740



741 **Figure 3. Further details for Bioregion 3.** (A) Lower interval estimate of probability of each site belonging to Bioregion 3. (B) Mean estimate
742 of probability. (C) Upper interval of probability. (D) The profile of species within Bioregion 3 -- species have been ordered according to their
743 overall prevalence (across all bioregions), each species' identity is preserved and can be used to understand the composition of each bioregion
744 (e.g. Hill et al. 2017). Results obtained by applying the methods in Foster et al., (2013) as implemented in Vanhatalo et al (in review).